

PADI (Preserving Access to Digital Information) and Safekeeping

Marian Hanley (*)

Abstract:

This report concentrates on the practical aspects of the National Library of Australia's PADI Safekeeping project, including selection, archiving and workflows. Some technical aspects of the National Library of Australia's in-house web archiving system, PANDAS, are also discussed.

Why Safekeep Resources?

Important digital information resources are in danger of becoming lost without good management. PADI is a subject gateway that attempts to bring together the most useful sources of advice and research relating to how digital information can be managed. However, identification is not enough if digital information is to remain available for use in the long-term. PADI Safekeeping not only seeks to identify but also to preserve key resources in its area of interest.

What is PADI?

PADI (Preserving Access to Digital Information) is a National Library of Australia initiative. The PADI web site is a subject gateway to digital preservation resources. PADI brings together a range of resources such as policies, project reports and journal articles covering a wide range of topics relating to the ongoing accessibility of digital information.

PADI is based on a model of cooperation. An International Advisory Group provides guidance for the PADI initiative. PADI currently works with two partners, DPC (Digital Preservation Coalition) and ERPANET (Electronic Resource Preservation and Access Network).

PADI extended the cooperative model by allowing registered contributors from around the world to add resources to the PADI database. Because many of these resources are in digital/electronic form and available only on the web, it was recognised that these resources are themselves in danger of being lost.

The Safekeeping Project

The commencement of the Safekeeping project further extended this cooperative model. The Safekeeping project aimed "to build a distributed and permanent collection of digital resources from the field of digital preservation" [1].


These resources are selected by a group of experts. They include "seminal papers which record a 'turning point' in thinking about digital preservation; or resources which define or describe an important issue, approach, project or study; or which summarise or raise important issues in digital

preservation". The PADI safekeeping selection guidelines are a refinement of the criteria used in selecting resources for inclusion in the PADI database. The refinements are based on type of resource, format and content. (see [Safekeeping Selection Guidelines](http://www.nla.gov.au/padi/safekeeping/archselcrit.html): <http://www.nla.gov.au/padi/safekeeping/archselcrit.html>).

The safekeeping program began in 2000/2001. Initially, it aimed to encourage resource owners to take appropriate action to archive their own resources. At this stage, the role of the National Library of Australia was to encourage and inform resource owners of best practice and to undertake safekeeping only if the resource owners could not archive their resource locally. This approach proved unsustainable because of the work-load involved in contacting and often recontacting resource owners. It became apparent that it would be more efficient to archive the selected resources ourselves, taking advantage of the National Library of Australia's technical infrastructure and expertise.

With the next round, the focus of the safekeeping program changed. In the paper *Safekeeping: a cooperative approach to building a digital preservation resources*, the authors state, "We are also keen to further explore, in our model of collaboration, possible 'natural' paths for transferring responsibility for preserving access for material for which owners cannot, or choose not to, assume a safekeeping role." [2]. The National Library was interested in seeing if, for example, parent institutions could be encouraged to take archiving responsibility for papers produced by their staff, or if regional or national libraries would accept identified resources into regional or national archives. When it came to practical implementation, however, we had to recognise that PADI itself was as "natural" an archiving partner as any other possibility, in that we had an interest in ensuring the chosen resources remained accessible because they were core to the business of PADI.

With this in mind, the process was revised so that the National Library could take a more direct approach to safekeeping. The safekeeping program was restructured so that resource owners were informed that the National Library was archiving their resource and that they had the option to reject the offer. All resource owners from the previous 2000/2001 safekeeping program were recontacted to inform them of the new approach. Almost 300 resources from the 2000/2001, 2001/2002 and 2002/2003 programs were identified, owners contacted and resources archived by the National Library of Australia.

On the PADI website, resources that have been selected as being of long-term interest or value and archived by the National Library of Australia are marked  .

In the next round of safekeeping it is hoped that procedures will be further streamlined so that archiving can take place at the time the resources are originally included in the PADI database. Resources would only need to be revisited if the Library received notification that permission to archive has not been granted.

The Safekeeping Process

There are four steps to the safekeeping process.

1. selection by nominated reviewers,
2. notification of resource owners,
3. archiving using the PANDORA Digital Archiving System (PANDAS), and
4. changing the status of the resource on PADI to indicate whether the resource was safekept, its Persistent identifier and date of archiving.

The 'Safekept' Selection Process

The selection process is currently based on a peer appraisal of resources on the PADI database. There are six selectors based in Australia, United States and Europe, including Norway and the United Kingdom.

Once the resources are selected and resource owners identified and given the option to decline the offer to archive their resource, the National Library uses the functionality of PANDAS for the archiving process. The selected resources are harvested via PANDAS, which is the system used to collect and manage web sites selected for long term archiving in the PANDORA archive, the National Collection of Australian Online Publications. Although the PANDAS methodology is used, international resources gathered as part of the safekeeping project differ from regular PANDORA titles in the way they are managed, including:

- As the resources are international they are separated from the National Collection of Australian Online Publications in storage.
- The resources are not listed, searchable or displayed via the regular PANDORA interface. Access is restricted until the 'live' resources are no longer available.

In other words, PADI 'safekept' resources are archived and managed similarly to the Australian resources chosen for inclusion in the PANDORA archive, but they form a separate bundle of resources.

The safekept resources remain identified through PADI. While a 'live' version remains available online from its original source, the PADI link continues to point to that version. If and when the resource becomes unavailable from that source, public access via PADI will switch to the archived version held by the National Library of Australia, with appropriate messages about viewing an archived version.

The process of harvesting and archiving the selected resources includes registering, gathering, quality control checking, re-gathering where necessary, applying metadata and applying access restrictions. A person was employed with funds provided by CLIR (Council on Library and Information Resources) for a period of 6 weeks to undertake these tasks.

Diagram 1 below illustrates the steps in the Safekeeping process flow.

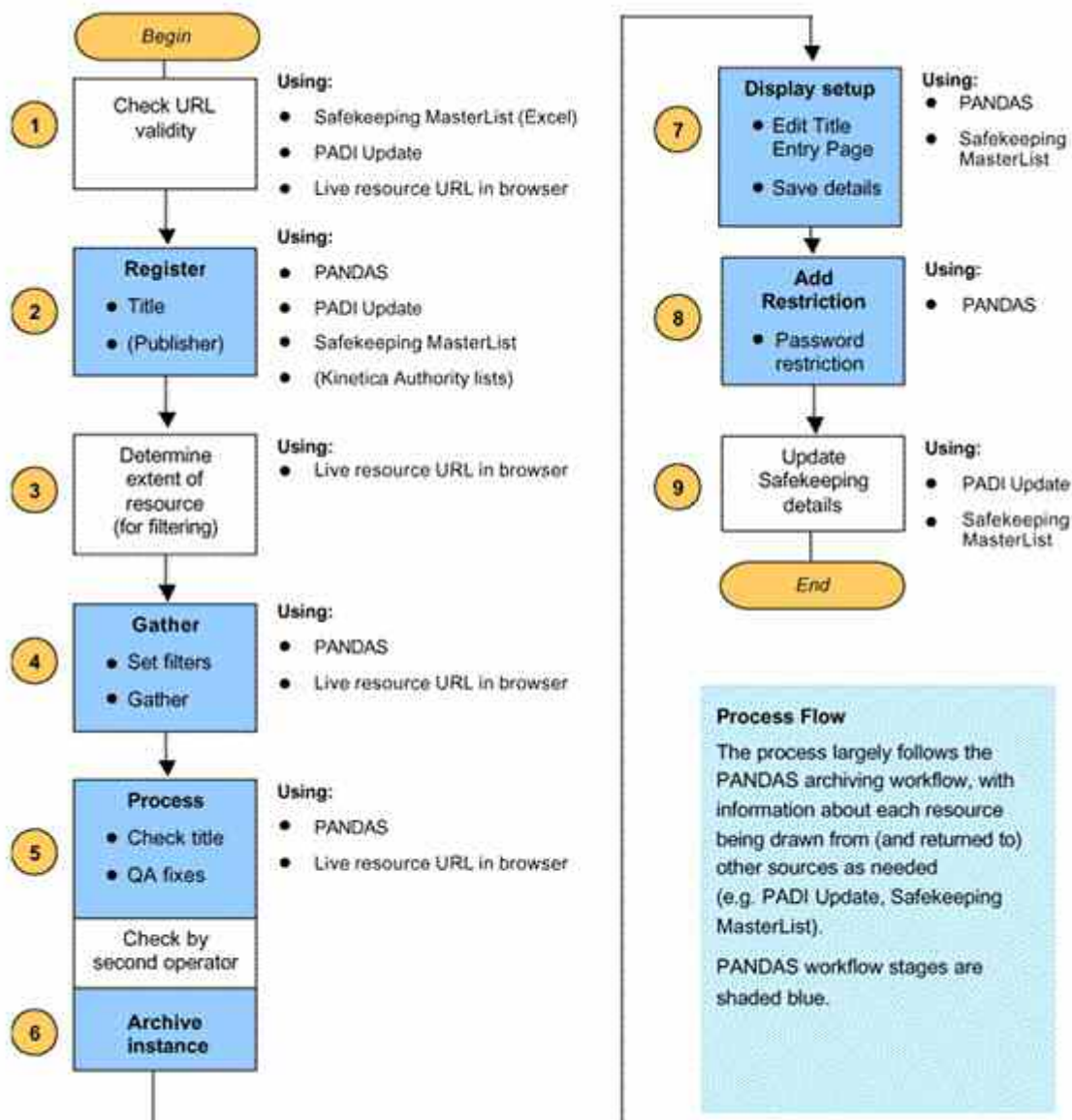


DIAGRAM 1: PANDAS Process Flow for PADI Safekeeping

How Archived Resources are Safekept

Technical Aspects

Archiving is undertaken using PANDAS (PANDORA Digital Archiving System), an archiving management system developed in-house by the National Library of Australia. It is the system used by the Library to gather resources for the PANDORA Archive (<http://pandora.nla.gov.au/>). PANDAS is a Web interface management system developed on the WebObjects platform (<http://www.apple.com/webobjects/>). It utilises available spider (crawler) software for the gathering of remote resources - currently the offline browser HTTrack (<http://www.httrack.com/>) is used. Gathered files are initially written to a "working space" server and the WebDAV protocol (<http://www.webdav.org/>) is incorporated into PANDAS to allow remote access to, and manipulation of, harvested files prior to being archived for preservation. The WebDAV protocol also allows files to be uploaded to PANDAS from local drives. Resources are archived for

preservation as TAR (Tape ARchive) packages on the Library's Digital Object Storage System (DOSS) which consists of a Sun E450 server, a CLARiiON FC4700 disk array, and a StorageTek Tape Library connected via a SAN switching infrastructure using LTO 2 data tapes.

PANDAS is used to record title and publisher metadata, to automate the scheduling of the gathering processes, and to manage access restrictions where applicable. MIME type metadata is automatically recorded for all files gathered from the Internet.

Three archived copies of a resource are maintained, two preservation copies on the DOSS (one an exact copy of the downloaded files and the other incorporating any changes made to the files after harvesting) and a copy which will be used for public access via the Library's Web server when the live version is no longer available. Multiple tape copies are made of all copies, in accordance with best backup practice and one copy is stored off-site.

Issues and Lessons Learnt

One of the major internal impacts of the project is in its time-consuming nature. The most time-consuming aspect of safekeeping is the process of determining the extent of the title to be gathered. Although entire sites are rarely gathered, time still must be spent deciding what groups of linked pages are relevant to the resource being gathered. Often a combination of filters must be devised to capture the essence of the resource. Decisions on what particular groups of linked pages are essential to the resource are based on whether the value or significance of the resource would be affected without the linked material.

Although the boundaries of the resource are often obvious, for example, from the directory structure itself, in a small number of instances this is problematic. The interconnecting nature of the web means that sometimes the resource boundaries are not clear and it is these resources that take disproportionately more time to archive.

As there has been a time delay of up to 12 months between adding resources to PADI and safekeeping them, some resources have disappeared from the web in any form by the time the safekeeping process searches for them. In other instances the resources may still be available, but the the URLs are altered. Determining the cause of broken links has been an accepted overhead, particularly as we have to be absolutely sure that a resource is no longer available before abandoning the safekeeping process.

Ideally, the person undertaking the archiving should have a cataloguing background, or at least an understanding of Authority Files, as publishers and names are registered within PANDAS according to AACRII rules.

Quality assurance is an important part of the safekeeping project. Resources are checked and re-checked at regular intervals throughout the process flow illustrated in Diagram 1. Resources containing numerous links, frames, plug-ins, or links to file types such as Powerpoint can be problematic to gather and often are identified as needing re-gathering and it is these resources that require the most checking.

Conclusions

The process of safekeeping needs to be done on a regular basis in order optimize the likelihood of success. Some of the resources identified earlier in the project are no longer extant on the web.

Although we attempted to contact them, many resource owners did not respond and it is assumed that these resources are now lost.

Taking an 'opt out' rather than 'opt in' approach to seeking permission to archive has resulted in positive response rates and has also removed the necessity for extending correspondence. There were no negative responses and many unsolicited positive responses to the National Library's request to archive in the latest round.

Although the safekeeping program has undergone extensive changes since its inception in 2000/2001, it has evolved so that all reasonable steps can be taken within an acceptable time frame to ensure that key digital preservation resources are not lost forever to the community.

References

[1] Berthon, Hilary, Thomas, Susan and Webb, Colin. Safekeeping: A Cooperative Approach to Building a Digital Preservation Resource. D-Lib Magazine, 2002. Available at <http://www.dlib.org/dlib/january02/berthon/01berthon.html>

[2] *ibid.*

Author Details

Marian Hanley

is the PADI administrator at the National Library of Australia.

Email: <http://library.cern.ch/HEPLW/9/papers/2/mhanley@nla.gov.au>

Acknowledgements

The author wishes to thanks Colin Webb, Kevin Bradley, Gerard Clifton and Paul Koerbin from the National Library of Australia for their contributions and feedback.